DOCUMENT RESUME

ED 094 098 UD 014 481

AUTHOR	Weber, George
TITLE	Uses and Abuses of Standardized Testing in the Schools. Occasional Papers, No. 22.
INSTITUTION	Council for Basic Education, Washington, D.C.
PUB DATE	May 74
NOTE	42p.
AVAILABLE FROM	Council for Basic Education, 725 15th Street, N.W., Washington, D.C. 20005 ($0.50)
EDRS PRICE	MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS	*Achievement Tests; *College Entrance Examinations; Criterion Referenced Tests; Cultural Factors; Educational Diagnosis; *Group Intelligence Tests; Multiple Choice Tests; Norm Referenced Tests; Reading Readiness Tests; *Standardized Tests; Test Bias; *Testing Problems; Test Reliability; Test Validity

ABSTRACT

It is asserted in this paper that some standardized tests do not do a good job of what they claim to do, and for some testing purposes nonstandardized tests are more appropriate or efficient. Under present circumstances, group I.Q. tests should be abolished. They provide no useful information that cannot be gained from achievement tests. And what they do provide--an I.Q. number for the cum card and the teacher's head--retards the academic progress of many children. On the whole reading readiness tests and the approach of which they are a part do more harm than good. It would be wiser to begin formal reading instruction, as some schools do, by attempting to teach all children the same things, without prejudicing or predicting their success. The standardized achievement tests given in the elementary and secondary schools are of little or no value to competent teachers in appraising the work of individual students. On the whole, college admission tests have a good record and continue to perform a useful function. If criterion-referenced tests tell us whether students have mastered instructional objectives, are they not better than norm-referenced tests? The answer depends on the purpose of the test. (Author/JM)

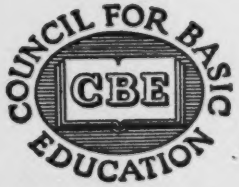
ED 094098

Uses and Abuses of Standardized Testing in the Schools

BEST COPY AVAILABLE

by

GEORGE WEBER
Associate Director
Council for Basic Education



UD014481

OCCASIONAL PAPERS
NUMBER TWENTY-TWO

A FEW HIGHLIGHTS FROM THIS PAPER

"Some standardized tests do not do a good job of what they claim to do, and for some testing purposes non-standardized tests are more appropriate or more efficient." (page 1)

"Under present circumstances, [group I.Q. tests] should be abolished. They provide no useful information that cannot be gained from achievement tests. And what they do provide—an I.Q. number for the cum card and the teacher's head—retards the academic progress of many children." (page 6)

"On the whole, . . . reading readiness tests and the approach of which they are a part do more harm than good. It would be wiser to begin formal reading instruction, as some schools do, by attempting to teach all children the same things, without prejudging or predicting their success." (page 9)

"The standardized achievement tests given in the elementary and secondary schools are of little or no value to competent teachers in appraising the work of individual students. The individual scores are of value to people outside the classroom, such as counselors, when they form a regular pattern over a number of years, and in occasional other circumstances. . . . Group scores can be of value to teachers in studying the relative progress of a class. Group scores, if they are interpreted and used with care, can be of value to others (principals, researchers, central office administrators, school boards, and the public) in making judgments about curriculum, programs, schools, and school systems." (page 21)

"On the whole, college admission tests have a good record and continue to perform a useful function. There are a few easily remedied flaws, and it is true that the scores have not always been used reasonably. But those who would abolish them would be hard put to offer good alternatives." (page 28)

"In order to discuss [the charge that standardized tests are culturally biased], cultural bias must be precisely defined. It is not enough to point out that the minority groups score lower, on the average, than does the white middle class. Bathroom scales are not biased against fat people because they record their obesity." (page 29)

"If criterion-referenced tests tell us whether students have mastered instructional objectives, are they not better than norm-referenced tests, which tell us merely how well a student does in relation to other students? The answer depends on the purposes of the test." (page 34)

Additional copies of this Occasional Paper may be purchased for 50 cents each, or 40 cents in quantities of twenty or more.

Printed May 1974

AUG 09 1974

BEST COPY AVAILABLE

CONTENTS

	Page
Introduction	1
Group I.Q. Tests	3
Reading Readiness Tests	7
Elementary and Secondary Achievement Tests	10
Some Faults of Reading Tests	
Interpreting the Scores	
Use of Group Scores	
The Value of Group Scores	
Summary	
College Admission Tests	22
Some Current Controversies	29
Are the Tests Culturally Biased?	
Is the Multiple-choice Format a Bad One?	
Are Criterion-referenced Tests Better?	
Suggestions for Further Reading	37

Uses and Abuses of Standardized Testing in the Schools

by

George Weber

INTRODUCTION

Hundreds of millions of test papers are written in our public schools every year. Most of these are teacher-made tests or tests provided by publishers for use with particular textbooks or other instructional materials. But a minority of these tests—a critical minority—are standardized tests.

A standardized test differs from other tests given by the schools in several ways. First, it is almost always constructed by test specialists. There is a recognized field of educational testing, and the people in it claim to be experts in the construction, interpretation, and criticism of standardized tests. There are a number of publishers of standardized tests for the schools, but only five or six account for most of the tests used. Secondly, a standardized test has explicit instructions for uniform (standard) administration. In most cases, the instructions are written on the assumption that the person giving the test has no experience in test administration and little or no knowledge of testing theory. Thirdly, a standardized test has norms (standards) for interpreting the test results, norms that have been derived from giving the test to a sample of persons presumed to represent the whole group of persons for whom the test is intended.

Some standardized tests do not do a good job of what they claim to do, and for some testing purposes non-standardized tests are more appropriate or more efficient. But standardized tests are used by all our public schools. Important and even critical conclusions and decisions are made on the basis of their results. For example, on the basis of their results the public is told that reading achievement is

going up or going down, an experimental program is deemed successful or unsuccessful, a child is placed in this or that class, and students gain admission to a particular college or fail to do so.

Standardized tests are relatively new. Group intelligence tests were developed during World War I. Today's elementary school achievement tests date from the first Stanford Achievement Test, published in 1923. The multiple-choice college admission tests started in 1926. But the big boom in standardized testing did not take place until after World War II. Today the typical schoolchild takes one to three such tests every year.

There are many kinds of standardized tests given in the school, in business, and in the armed forces: intelligence tests, academic aptitude tests, achievement tests, personality tests, attitude tests (or inventories, as they are usually called), interest inventories, and vocational aptitude tests. This paper will deal only with the first three categories. Personality tests, attitude inventories, interest inventories, and vocational aptitude tests are less widely used in the schools, more controversial, and differ considerably from intelligence, academic aptitude, and achievement tests. The next four sections will describe the tests and discuss their uses, limitations, and abuses. The last section will deal with some of the current controversies surrounding these tests.

The Council for Basic Education hopes that this paper will contribute to the understanding of this important school activity. We have tried to use terms and concepts that will be clear to the nonspecialist.

GROUP I.Q. TESTS

Most of our public schools maintain, for each pupil, a cumulative record card, often simply called a "cum card." It contains a lot of basic information about the child, including name, date of birth, class assignment, and the scores the child has received on standardized tests. Usually there is a place for an I.Q. number. The figures range from 130 and above ("very bright") down through 110-130 ("bright"), 90-110 ("average" or "normal"), 70-80 ("dull"), to below 70 ("feeble-minded"). This magic number of two or three digits has been obtained, in almost all instances, from the administration of a group I.Q. test, that is, a paper-and-pencil, standardized, multiple-choice I.Q. test given to groups of children.[1]

The most widely used group I.Q. tests are probably the *Otis-Lennon Mental Ability Test* (published by Harcourt Brace Jovanovich), the *California Test of Mental Maturity* (California Test Bureau), the *Otis Quick-Scoring Mental Ability Test* (Harcourt Brace Jovanovich), the *Lorge-Thorndike Intelligence Test* (Houghton Mifflin), the *Henmon-Nelson Test of Mental Ability* (Houghton Mifflin), and the *Kuhlman-Anderson Intelligence Test* (Personnel Press).

Although these group I.Q. tests have been under attack in recent years and have been banned by the New York City and Washington, D. C. schools, they are still widely used, misused, and misunderstood.

A standardized test is supposed to have a high degree of what test specialists call "reliability" and "validity." Reliability means consistency of result. If a reliable test were taken again by the same person after a brief period, his score would be very close to what it was the first time. Validity means that the test measures what it is supposed to measure. For example, a college admission test, to be valid, would have to predict success in college very well. Group I.Q.

[1] These group I.Q. tests should not be confused with individually administered I.Q. tests—notably the Stanford-Binet and the Wechsler. The individually administered tests have some of the shortcomings of the group tests, but in the hands of a skilled psychologist they can sometimes be of value. The psychologist should be able to insure that directions are understood and that the person being tested is well. Moreover, he should be able to take into account the test-taker's background in evaluating the results. Individual tests tend to have more nonverbal items, but they are still culture-bound in the broad sense.

tests are reliable measures of something, but for the reason explained below they are not valid measures of intelligence.[2]

The theory behind the group I.Q. test is quite simple. The same questions are given to a large group of people of the same age. Those who get higher scores must be more intelligent than those who get lower scores. The implicit assumption is that all of the persons being tested have had substantially equal opportunity to learn the types of things included in the test. That is just not the case unless the tasks are entirely new to all. (Finding completely new tasks would be difficult in any event; as shown below, these tests do not even seem to try.)

The group I.Q. tests are heavily dependent on verbal knowledge. On one of the above tests, for example, second- and third-graders are asked questions which cannot be answered correctly without knowing what the following words mean: *mechanic, each, hinge, funnel, banister, crouch, duet, portable, silhouette.* On another one of the above tests, third- and fourth-graders are asked which one of the following words does not belong to the group: *general, major, private, colonel, captain.* And which of these does not belong: *foolhardy, cautious, reckless, daring, rash.* Another of the tests asks pupils in grades 3-6 questions which depend on knowing the following words: *tiny, birch, velvet, tour, liquid, insinuate, miserly.* For example, one question reads: "Velvet is a kind of: (1) deed (2) cloth (3) brush (4) picture (5) leather." Obviously tests so highly dependent on particular verbal knowledge are not valid measures of intelligence.

Since these tests have high reliability (the big test publishers are good on reliability but often poor on validity) what *is* it that they are measuring? Although some of the tests have sections which involve tasks which may well be new to all those tested, over-all each test contains a heavy component of verbal knowledge, general information and, in some cases, mathematical knowledge. What these tests are measuring, then, is a kind of general academic achievement in standard English.

---

[2] There is a difference between *intelligence* and *I.Q.* Although there are many definitions of intelligence, they all involve the notion of general mental ability — the ability to understand, the ability to learn, the ability to adapt, the ability to solve real problems. I.Q. is the measure of intelligence in terms of age. Since up to a certain point, usually regarded to be adolescence or early adulthood, intelligence normally increases, I.Q. relates this intelligence (mental age) to chronological age. For example, a child of eight who has the intelligence of an average eight-year-old would have an I.Q. of 100 (eight, his mental age, divided by eight, his chronological age).

In fairness to the publishers, the "fine print" of the administration manuals sometimes points this out and cautions against the conclusion that the tests measure an inherent or constant mental ability. But these clauses have about as much value as the phrase "If pain persists, see your doctor" on a widely advertised nostrum. Back in college, our teachers, counselors, and principals learned about I.Q., the significance of the various numbers, and the bell-shaped curve that shows the distribution of scores: most in the middle and a few at each end. It is not surprising that tests called intelligence tests that are published by well-known firms, and that provide an I.Q. number in the usual scale, are widely regarded as valid measures of intelligence.

For what are these tests used? To predict academic success, to group pupils for instruction, to select materials and learning tasks, to determine speed of instruction—these are some of the purposes suggested by the publishers. Grouping pupils for instruction on the basis of what they already know and can do is a basically sound practice. (It has recently become fashionable in the form of "individualized instruction," which may or may not involve grouping.) And if pupils are grouped for different instruction, they should have appropriate materials and tasks—appropriate to take them from their present level of skill and knowledge to a higher level. These purposes are best served by the use of careful measures of specific achievement. But a deliberately slower speed of instruction may well help to fulfill the prophecy of lower academic success. And those I.Q. numbers attach a label to a child ("bright," "average," or "dull") on the basis of an invalid test. It is not surprising that the children in the slow group wind up learning least.

Another use of the I.Q. score derived from these tests is to label children "underachievers" and "overachievers." Both groups are regarded somewhat as if they were diseased. An "underachiever" is a child who is not living up to the potential supposedly indicated by his I.Q. score. An "overachiever" is a child who is exceeding the potential supposedly indicated by his I.Q. score. Exactly how he can do this is not clear: "overachiever" is one of the strangest words in the often strange jargon of education. Both concepts rest on the unfounded assumption that the group I.Q. test has provided a better index of the child's ability than anything else, including his demonstrated day-to-day achievement. We would be better off without these terms.

Collective scores from I.Q. tests are sometimes used as an index of the "potential" of the group (class, school, school system), with

which is compared its "achievement," as measured by an achievement test. Generally speaking, this procedure is fallacious because it amounts to comparing the result on one general achievement test with that on another. The fact that the results are somewhat different is not surprising: results of two achievement tests usually differ, too.

Summing up, group I.Q. tests might be occasionally useful as tests of general school achievement if they did not produce I.Q. numbers and if they were called "general achievement tests." Under present circumstances, they should be abolished. They provide no useful information that cannot be gained from achievement tests. And what they do provide—an I.Q. number for the cum card and the teacher's head—retards the academic progress of many children.

READING READINESS TESTS

A very specialized kind of standardized test, the reading readiness test is a very important one. It is given to most children, normally at the end of kindergarten or at the beginning of first grade, for the purpose of determining their readiness for formal reading instruction. Its importance lies in the fact that on the basis of the child's score he is put in a group that will be given reading instruction at a normal pace, put in a "slow group," or not allowed to begin formal reading instruction at all.

The concept that a certain mental and physical development is desirable, even necessary, before reading instruction is begun is beyond question. Obviously, reading instruction cannot begin the day after a child's birth. Toward the other extreme, for about four decades it has been conventional wisdom among school people in this country that a "mental age" of six and a half years is the desirable time to begin formal reading instruction. (This dictum dates from the research published in 1931 by Morphett and Washburne.) Since most schools begin formal reading instruction in the first grade, the practical result has been that some children, deemed "ready," have begun reading at a standard pace at the beginning of first grade. Others, deemed not yet ready, have begun at a slower pace or, more likely, have been given instead a pre-reading or reading readiness program for some or even all of their first-grade year.

One method of determining readiness has been to ascertain the "mental age" of the child, that is, his chronological age adjusted by his I.Q. score. (The "mental age" of 6½ is that of a 6½-year-old child with average I.Q., a younger child with a higher I.Q., or an older child with a less-than-average I.Q.) A more common method of determining readiness is to give a standardized reading readiness test.

Four widely used reading readiness tests are the *Murphy-Durrell Reading Readiness Analysis* (published by Harcourt Brace Jovanovich), the *Gates-MacGinitie Reading Skills Test* (Teachers College Press), the *Lee-Clark Reading Readiness Test* (California Test Bureau) and the *Metropolitan Readiness Test* (Harcourt Brace Jovanovich).

These tests are orally administered to groups. Largely untimed, they vary in length from 20 minutes to about two hours. The Murphy-

Durrell test involves first teaching some letter sounds and then testing them (24 items); testing knowledge of letter names (52 items); and first teaching and then testing whole words (18 items). The other three tests are basically different in that no teaching is done. They test several of the following: listening vocabulary; the ability to pick the one word or letter in a set of four that is different from the others; the ability to copy a letter, a number, or a figure; knowledge of the alphabet; the ability to match up letters; general information; and knowledge of number names and very simple arithmetic.

The pictures, words, and concepts of these tests are biased in favor of children from educated families. (The bias is smallest in the case of Murphy-Durrell, which gives a child some opportunity to learn the material during the test itself.) The use of one of the tests accordingly often results in delaying reading instruction for the very children who are already behind. Then, because they have been taught less in the first grade, they are behind in the second grade and probably never catch up to their contemporaries who were judged on the basis of one test to be more ready to begin reading instruction. Some of the manuals state that the test should not be the "sole measure" of reading readiness, but this qualification cannot be taken very seriously when the test presents a clear basis for decision and the other factors to be considered are not spelled out. One manual baldly states the usefulness the test can have in defending the school's policy of delaying instruction: "It is often . . . difficult to tell parents that 'in my *judgment*' the child is not ready to begin reading. It is relatively easy, however, to say that a child has been given a test, the results of which indicate placement. . . ."

Another flaw in the tests is the weight they give to some simple knowledge (for example, the names of the letters) which could be taught in a very brief period to the children who do not know it. (This weight given to letter names is paradoxical since many reading specialists discourage teaching the letters to pre-school children at home.)

These reading readiness tests rest on two notions that have come under increasing attack: that formal reading instruction should begin no earlier than the start of the first grade, if then; and that reading instruction for the disadvantaged should be delayed. In contrast to the first idea, in recent years many schools have put beginning reading in the kindergarten, and there are suggestions that it begin even earlier, at least for some children. With respect to the disadvantaged, most recent thinking advocates beginning their formal instruction

earlier than for their more knowledgeable contemporaries (see, for example, Head Start[3] and the Bereiter-Engelmann program[4]).

On the whole, these reading readiness tests and the approach of which they are a part do more harm than good. It would be wiser to begin formal reading instruction, as some schools do, by attempting to teach all children the same things, without prejudging or predicting their success. Then, after perhaps a half-year or a year, the pupils could be grouped for further instruction on the basis of their demonstrated achievement. There would be two advantages to such a procedure. First, it would give all children a better chance to respond to instruction and to allow factors other than those measured by a reading readiness test to come into play. Secondly, it would shift the task of evaluation from a single test to the teacher, which is where it ought to be.

---

[3] Head Start is the Federally financed program which aims at teaching disadvantaged four- and five-year-olds many of the things that other children learn at home.

[4] The Bereiter-Engelmann program, published as DISTAR by Science Research Associates, involves very formal language instruction for disadvantaged pre-school and primary-grade children.

ELEMENTARY AND SECONDARY ACHIEVEMENT TESTS

Most of the standardized tests that children take in school are achievement tests. While over the thirteen years of elementary and secondary school (counting kindergarten) a child may take several I.Q. tests, one reading readiness test, and several college admission tests (discussed in the next section), he will almost certainly take six or more achievement tests, and he may well take as many as eleven or twelve. Some school systems are now giving achievement tests in every grade, or in almost every grade. Six widely used achievement series are the *SRA Assessment Survey* and the *Iowa Tests of Educational Development* (published by Science Research Associates for grades 1-12); the *Stanford Achievement Tests* (Harcourt Brace Jovanovich, grades 1-9); the *Cooperative Primary Tests* (Educational Testing Service, grades 1-3); the *Iowa Tests of Basic Skills* (Houghton Mifflin, grades 1-9); the *Metropolitan Achievement Tests* (Harcourt Brace Jovanovich, grades 1-6); and the *Sequential Tests of Educational Progress* (Educational Testing Service, grades 3-14). What do such tests measure? How well do they do it? For what purposes are the results used?

All of the six test series named above purport to measure reading and mathematics achievement at every level. Beyond the primary grades some or all purport to measure, in addition, achievement in science, social science, punctuation, English expression, and study skills. While these are very important subjects, it should be borne in mind, in view of the tests' general names and the tendency to equate their scores with general academic achievement, that not one of these tests tries to measure directly either of two important skills: speaking and writing. Nor do they attempt to test many other things that many people believe are important outcomes of good schooling: for example, interest in learning, initiative, imagination, morality, self-discipline, knowledge and skills in the arts, ability in foreign languages, and physical development. Turning to the form of the tests, they are all multiple-choice tests. All except the Cooperative Primary Tests are timed: that is, they are given with specific time limits with the knowledge that many or even all of the students will fail to finish within the indicated periods. (The lowest level of the SRA Primary I, is a

partial exception since it has a time limit so generous, according to the publisher, that about 90 per cent of the pupils finish.) The series are arranged in different levels, each one designed for a different grade range. The content of each is supposedly related to the usual curriculum in those grades. The Stanford series, for example, has six levels, designed respectively for grades 1-2, 2-3, 3-4, 4-5, 5-6, and 7-9.

How well do these tests measure what they claim to? The answer depends upon the purposes for which the results are used, the level of the test, the subject matter involved, and the competence and integrity of their administration.

The purposes can be divided between those related to the scores for the individual students and those related to the scores for groups. The individual's scores are used to promote to the next grade, to assign to a class within a grade or to a group within a class, to assign to a special program (for example, for the retarded or the gifted), and to indicate the kind of courses he is permitted to take in junior high school and high school. They are also used, sometimes, to indicate the areas of his schooling that need greater attention. The individual's scores are also used in a counseling function, that is, in talking with the student and his parents in connection with choices in which they at least nominally take part: courses, programs, need for outside help, career plans, and so forth. Finally, and sometimes subtly, the individual's scores may well influence the teachers' expectations about the student.

Generally speaking, these purposes are somewhat better served by these tests the higher the grade level of the test and are more poorly served in reading than in other areas. Over-all, the tests provide less information about the student's achievements in the various fields than is already known to his teachers, assuming them to be even moderately competent. For example, the classroom elementary school teacher should know more about her pupils' individual reading and mathematical ability, either from their previous teachers or from her own experience, than the tests can tell her, but she has been told that her judgments are not as dependable as scores on standardized tests. And when it comes to particular skills within these areas, the tests are even less helpful. For example, an arithmetic test may include a half-dozen multiple-choice questions on long division. There is no way of knowing how many of these are answered correctly by chance. A teacher can get a more reliable index of the child's skill by giving him six problems to do in the usual way (no guessing possible).

Some Faults of Reading Tests

The tests in reading deserve special attention because they are usually given the most weight in the various decisions listed above. One of the faults of the reading tests at the lower levels is the frequency of inappropriate vocabulary. The test makers say that the words used are from the children's hearing and speaking vocabularies, but most second-graders do not know *log, chimney, tale, ribbon, village, polite, chatting, sapling, postage, baggage, peak,* and *harvest.* By using a number of words such as these, the test makers have made a reading test dependent in part on breadth of vocabulary. Breadth of vocabulary is one aspect of reading skill in the higher grades and in high school, but in testing the mechanical skills of beginning reading it penalizes the child with a small hearing vocabulary.

Another important fault of these tests is their cultural bias. All of our children, including those from inner-city homes and from homes where Spanish or some other foreign language is spoken, should learn to read standard English, and it is appropriate that tests should try to determine their ability to do so. But selection of vocabulary, pictures, and subject matter should not bias the tests in favor of some groups of children, and in some cases it does. Cutting the grass, raking leaves, going to the beach, taking a vacation trip in a car, having a pet turtle, and celebrating birthdays with a party for other children are more likely to be familiar to some groups of children than to others. In the lower grades, the solution to the vocabulary faults outlined above lies in using words that are familiar to all English-speaking American children of the particular grade level. For organizations as large and as talented as the big test makers this should present no difficulty, since the number of words understood by ear by all of these children in the lower grades runs into the thousands. Pictures and subject matter present a somewhat more difficult problem, but a great deal can be done to make the tests fairer in this regard. Incidentally, the newest tests are less culturally biased than those of a decade or two ago, but much improvement remains to be accomplished.

Another fault of the reading tests lies in the structure of some of the reading comprehension questions. In theory, these questions test the child's ability to read a paragraph, but in fact some of the questions can be answered correctly without even reading the paragraph.[5]

[5] The frequency of this flaw is analyzed in "Determining the Passage Dependency of Comprehension Questions in 5 Major Tests" by J. Jaap Tuinman (*Reading Research Quarterly*, IX [1973-1974], 2).

Moreover, a few questions cannot be answered correctly without information that the paragraph does not provide. In such cases, the test is measuring the extent of general information known, not reading ability. Here is an example of a question that can be answered correctly without reading the accompanying paragraph: "At the museum there were ( )paintings ( )books ( )dogs." And here is a question that requires information not provided in a paragraph about a boy who comes from a country where almost everyone speaks Spanish: "Juan might have come from ( )Germany ( )England ( )Mexico ( )France." The paragraph makes no mention of Mexico, and so the pupil has to know that Mexico is a Spanish-speaking country. These are faults that are easily remedied. In fact, most of the reading comprehension questions *do* require reading of the paragraph and can be answered correctly with the information it contains. There is no reason why they all could not meet these obvious requirements.

Still another fault of the reading tests (and one that applies to other tests as well) involves their time limits and their scoring. The scoring on all these tests is based on the number of items right, with no penalty for wrong answers. Since, with the exception of the Cooperative Primary Tests, these are all timed tests, some or even all of the students will fail to finish. Yet a pupil who gets 30 right and 10 wrong on a 40-item test gets exactly the same score as a pupil who does only 30 and gets all of them right! Obviously this gives an advantage to the pupil who works faster because of habit, coaching, or greater familiarity with this kind of test. This fault could, and should, be eliminated or reduced in significance. There is no reason whatsoever why the tests cannot be scored in a way that penalizes wrong answers. And the time factor could be either eliminated by allowing all children to finish (as in the case of the Cooperative Primary Tests) or greatly reduced in importance by extending the time allowed.

These faults greatly limit the value of individual scores, even for the teacher who is incompetent or lacks confidence in her own judgments of her pupils' abilities.

Interpreting the Scores

Beyond the tests themselves and the raw scores that result, there are weaknesses in the testing procedure that relate to the method of interpreting the scores. The raw scores are typically interpreted in terms of national norms, which are estimates of nation-wide perform-

ance. These national norms are usually recorded and reported as grade-equivalents. In other words, the student's score is recorded as 4.7 (fourth grade, seventh month) in reading, for example, or 8.6 (eighth grade, sixth month) in social studies. This method of interpreting the student's achievement is so important that it requires further explanation.

The national norms are derived from giving the test to what is supposed to be a representative sample of students. But since the samples are different and are taken at different times, the norms for different tests vary. As a result, the normed score for a student depends partly on which test he takes. Recognizing this fact, in 1971 the U.S. Office of Education granted some $700,000 to the Educational Testing Service for a project to develop conversion tables for the various national norms on the reading and vocabulary sections of seven tests widely given to grades four, five, and six.

The method of reporting national norms in terms of grade-equivalents is one which Dr. Henry S. Dyer, a former vice-president of Educational Testing Service, has called "absurd, wrong and misleading."[6] The grade-equivalent represents the estimated median score that pupils in that month of that grade would achieve on the test nation-wide. For example, a 3.8 in reading would mean the average score for a child in the eighth month of the third grade. One trouble with this procedure has already been referred to—the fact that the grade-equivalent score for a given person varies from test to test. Moreover, on some tests, a few answers one way or the other can make as much as a whole year's difference in the grade-equivalent score. Obviously, the tests are not that accurate. Another problem lies in the frequent misconception that the score means that a child has mastered the standard curriculum up to that point in schooling: that is, that the child with the 3.8 score has mastered reading through the material given in the eighth month of the third grade. The score has no such meaning. Even if the 3.8 grade-equivalent were always an accurate estimate of the average achievement for the child in the eighth month of the third grade, that average child has not *mastered* the reading curriculum to that point: he is merely at that point in his schooling.

Parents, and even teachers, usually do not understand this. If Susan, in the third grade, gets a 6.8 score on her reading test, her proud parents usually believe that she reads as well as typical children in

[6] *The New York Times*, March 23, 1971.

the eighth month of the sixth grade. This is usually not the case. Even if the norming is accurate, it means that Susan, *on a third-grade test,* got a score equal to the average score that children in the eighth month of the sixth grade would get *on that test.* If Susan took a reading test intended for sixth-graders, with much more difficult vocabulary, syntax, and content, she probably would not receive a 6.8 *on that test.* It is important to understand the distortions at the other end as well. Johnny, a fourth-grader who takes the SRA reading comprehension test intended for fourth-graders and receives a 2.6 score, is, in all probability, not reading as well as the average pupil in the sixth month of the second grade in any useful sense. That score is the median chance score on that multiple-choice test if all questions are answered, and so Johnny, quite possibly, is not able to read at all! No wonder Dr. Dyer had such harsh words for grade-equivalent scores.

Fortunately, most tests offer national norm scores in other terms. (In fact, the Cooperative Primary and Sequential tests do not provide grade-equivalent tables.) A better way of interpreting the raw score is in terms of percentiles. In this method, a 35 means that the pupil's score is estimated to be better than 35 per cent of children at his stage of schooling. The weakness of this scoring system is that it conveys a false sense of accuracy. The child who scores 35 may well score 25 or 45 on another test or at another sitting. Because of this problem, some tests offer stanine or percentile-band scores. Without going into the details, it can be said that these scores indicate broad ranges of achievement. In the stanine system, the score ranges from 1 (very low) to 9 (very high). These rough-cut systems reflect the rough accuracy of the tests themselves and avoid the many problems of grade-equivalencies. They are adequate for reporting to parents and are less subject to misinterpretation by the teachers who want to use them.

Besides teachers and parents, others use the individual's scores on the standardized achievement tests, notably counselors, principals, and the system as an institution. The scores, rather than teacher judgments, are often used for the purposes listed earlier: promotion, assignment, and counseling. If all our teachers were competent and confident of their judgments, this would be unnecessary. But in the real world, many teachers are incompetent or lack confidence in their judgments, and the test scores seem to offer a disinterested measure of the individual's achievements. As already mentioned, the measure is actually a partial one and one that is deficient in other

respects as well. Nevertheless, if the scores form a consistent pattern over the years, they may be of help to the principal or counselor in some situations, particularly in placement of students coming in from other schools. In recent years, authorities have used the scores in an attempt to change promotion practices. In New York City, for example, the chancellor (superintendent) recently ordered that a student would henceforth not be given an eighth-grade diploma if he were more than a year behind in reading on the basis of the standardized test. The student so behind will, however, be given a certificate that will enable him to proceed to high school, or be "transferred" with neither diploma nor certificate. This order tightened up an effort started several years ago to eliminate the embarrassment of turning out high school graduates who are functional illiterates. This shameful situation illustrates the occasional usefulness of the standardized test scores in circumstances (by no means confined to New York City) where existing school systems leave much to be desired.

Uses of Group Scores

Group scores on achievement tests (scores for classes, other groups of students, schools and whole school systems) serve purposes quite different from those served by individual scores. Group scores are used by the teacher to judge the progress of the class in the several skills, by various people to assess curriculum, and to evaluate projects, teachers, schools, and school systems. Here again, the grade-equivalent national norms are usually used, despite the shortcomings stated above.

For the first purpose, the teacher can rely on the class scores in the various skills and subjects with somewhat more validity than she can on the individual scores. The reason for this is statistical: while a given pupil's score on a few long division questions is subject to considerable chance, the combined score of 25 pupils is subject to a great deal less. The teacher can therefore use the relative scores to compare achievement in long division and, say, in multiplication. She might conclude that she needs to give more attention to the skill with the lower score. The same conclusion, however, can be reached by the competent teacher by other means, notably tests of her own making. The situation is similar when the test covers a number of other subjects. Results on the tests that cover several subjects may suggest that the class is doing relatively worse in science than in, say, reading, and the teacher can adjust her teaching accordingly.

The other purposes of group scores involve a fundamentally different situation—that of scores being used by people outside the classroom. The principal and assistant principals of a school, particularly a small school, in appraising the pupils' collective achievements, the effectiveness of various materials and programs, and their teachers' competence, should be able to make better judgments than those based simply on the group scores of the various classes and grades. Yet the group scores may be of some value if the principal or assistant principal understands the tests and knows the background of the pupils. When it comes to people physically outside the school—the researcher, central office administrators, the school board, parents, and the public—there is no convenient way to evaluate the students' collective achievements except by the group scores on the tests. In these cases, the tests, despite their shortcomings and abuses, provide the best information available. For reasons that will be outlined below, the scores must be interpreted with great care.

The fundamental weakness of the use of group scores lies in the fact that one cannot always count on honest and competent administration of the tests. Typically, the tests are given by teachers, but occasionally they are given by project directors or researchers. The accuracy of the scores depends on precise administration. It also depends on honest administration, and as more importance is being attached to the results, the temptations to cheat are becoming greater. It is well to remember that the tests are being administered, in most cases, by persons who have, or think they have, something at stake in the outcome. A teacher wants to be judged successful; a project director is in the same position; and a firm with a performance contract has perhaps most of all at stake. The brief experience with performance contractors has already led to efforts to provide independent, outside auditors. Most teachers and administrators try to administer the tests honestly. But the possibility of cheating must always be borne in mind, particularly in the lower grades, where the nature of the tests makes cheating quite easy.

Most cheating takes the form of coaching (excessive preparation), but there are other, more flagrant types of misbehavior, such as teaching the particular words or items that appear on the test, practicing on the test itself, changing the answers before the tests are scored, giving pupils aid during the test, allowing additional time, and failing to test pupils who are expected to do poorly. Since the tests do not penalize wrong answers and most have time limits, even such apparently innocuous acts as urging the children to guess or

urging them to work more rapidly can make a significant difference in the scores. It is no wonder that some schools have considerably higher test scores than their students' day-to-day achievement warrants.

The Value of Group Scores

Taking into account this problem of competent and honest administration, what can we say about the value of group scores for various purposes? They are sometimes of value in assessing curriculum. The state of California saw its elementary test scores in mathematical computation decline year after year for seven or eight years coincident with the state's adoption and increased use of a "new math" series of textbooks. Although many of the teachers in the state were undoubtedly aware of the decline in computational skills in their own schools, the decline of the scores state-wide, at the same time as reading scores were *not* declining, was a fairly conclusive way of locating the difficulty. The same thing happened in the state of New York. Without widespread routine achievement testing there probably would have been great argument about whether computational skills were declining, and, if so, whether the "new math" programs were responsible.

The two cases point up a number of things. The routine administration of achievement tests provided data that could later be analyzed in various ways. If the tests had been administered on an *ad hoc* basis to determine the accuracy of suspicions about the impact of the "new math" on computation, the testing and investigation would have taken years, which would have delayed conclusions and remedies. Moreover, a special testing program would have been subject to great argument over the proper tests to give and the procedure, and perhaps even the integrity of the testing would have been jeopardized. The valuable data were available in a timely fashion because they resulted from a-testing program that was routine and done without previous knowledge of the purpose to which it was to be put.

Similar applications of routine test results can be made in other fields. For example, a school system may find that its scores in English usage are far below those of other systems in similar localities, or far below their own scores in other subject areas. Such a finding can properly lead to an investigation of its curriculum. Note that I have mentioned two different kinds of analyses: one that compares test results over time and one that compares test results at a single point in time. Both can be useful.

Using test scores to evaluate a particular project, program, or ex-

periment is difficult business. Even if the test is appropriate to the purposes of the particular project, the testers have a stake in the outcome; they are the people who have advocated the experiment or who publish the materials being tried out. An independent evaluation is necessary, but even that is not as easy to achieve as may first appear. If those who have a stake in the experiment know the test that will be used to evaluate it, it becomes possible to "teach to the test" and thereby improve the chances of a favorable evaluation. It is desirable to allow the independent evaluator to use any reasonable test he chooses, without informing the interested parties of the choice in advance. Even that is not as sure a guarantee as it may seem because of the similarity among the various standardized achievement tests.

An additional safeguard for sound evaluation is to provide for an experimental period of at least three years. As Dr. Roger Lennon, executive vice-president of Harcourt Brace Jovanovich, has pointed out, the standardized achievement tests are not accurate enough to form the basis for judging experiments when the period is a year or less.[7] We must therefore question the marvelous success stories about programs that raised the scores of the pupils so many months during an instructional period of so many months. On experimental programs, we must demand an independent evaluation that shows, over a period of three years or longer, a gain in achievement of a year or more *over what could otherwise have been expected.* There are such success stories, but they are few and far between.

The use of achievement test scores to evaluate the performance of teachers is theoretically possible, perhaps, but not now practical. One obvious flaw is that the teacher to be rated is giving the test herself! Another obvious difficulty lies in the fact that most teachers have their students only one year. When a fairly discrete subject is taught in one year, as, say, in the case of high school physics, the quality of the teaching might make a substantial difference in the year-end test scores. But in most subjects—reading, mathematics, social studies—the teacher of one year is building on what others have done, and the subjects are too extensive to be mastered in a year. Accordingly, a judgment as to what a teacher has achieved during the year should take into account not just the level of achievement at the year's end, but the level of achievement at the beginning of the year and the learning ability of the students she had to teach. A competent prin-

[7] "Accountability and Performance Contracting," an address to the American Educational Research Association, New York City, February 5, 1971.

cipal knows the ability of the teachers in his school, and this judgment depends only in part on the achievement test scores of the pupils.

For some years, almost all school systems have been compiling achievement test scores school by school. Beginning in 1966 with the publication of scores for the New York City schools, a small number of school systems have released these results to the public. Unfortunately, these scores cannot always be taken at face value. School personnel know that some schools have cheated on their tests, but the scores have to be published anyway. Even though perhaps 90 per cent of the schools have roughly accurate scores, there is usually no way for the outsider to know, short of conducting an independent evaluation, whether a given school's scores are sound. One can safely assume that where the scores reflect average or below-average performance in relation to the background of the students, the results are roughly accurate. It is where the school reports high scores in relation to the background of its students that skepticism is in order. But these scores do provide useful information to the central administration and the school board, and they give the general public a better picture of academic achievement than it would otherwise have.

The users of these figures should, however, bear in mind a number of things. The most important, perhaps, is that an average score means true accomplishment for a school where the background of the students is poor, whereas it means the opposite for a school where the background of the students is good. That is, the quality of the school's effort cannot be inferred simply from its test scores. Another fact is that the scores are usually reported only as a single-figure average, which does not show how many students score far above or far below that point. Another fact is that all of these scores are in terms of national average achievement, and there is no indication whether this is good, bad, or indifferent in terms of reasonable standards. It may seem strange, but the fact that a given school's third grade does as well on a given test as the national third-grade average does not reveal how well the children can read. Since reading achievement in the primary grades is generally below what could reasonably be accomplished, reading scores suggest a better achievement than is in fact the case. For example, study the scores as long as you like and you will still not learn that from 10 to 15 per cent of our third-graders, nation-wide, cannot read at all.

A fair comparison of achievement scores between school systems is difficult and often impossible. One reason is that different school systems use different achievement tests, and the norms are not easily

comparable. Another reason is that the students may not be similar. For example, if one wanted to compare achievement scores for Baltimore with those of Chicago, one would discover that Baltimore uses the Iowa Tests of Basic Skills and Chicago uses the Metropolitan. Besides, although Baltimore and Chicago may seem to have similar groups of students, one would have to go into this question carefully to determine if this were actually true.

While fine comparisons between school systems are very risky, gross comparisons are possible. If one system's sixth-graders have an 8.4 median score in reading and another system's have a 6.4, one can safely assume that reading achievement at the sixth-grade level in the first system is higher, even though different standardized tests were used. If the same test was used, a narrower difference will allow the same conclusion. It should be remembered, however, that these are single-figure averages. The highest *individual* scorers in the 6.4 system will be above the lowest scorers in the 8.4 system. Usually the highest-scoring *school* in the 6.4 district will have a higher score than the lowest-scoring school in the 8.4 district, though in a given case this may not be so because the 6.4 district may be a large city in which the children from the most favorable backgrounds in the metropolitan area are in the suburban districts and private schools and the 8.4 district may be a homogeneous suburb of families with highly favorable backgrounds. As a general rule, schools differ in achievement more than do school systems, and individual children differ most of all. In a study of inner-city schools that I made several years ago, the school with the lowest over-all reading achievement at the third-grade level nevertheless had 19 per cent of its third-graders scoring above the third-grade level.

Summary

In summary, the standardized achievement tests given in the elementary and secondary schools are of little or no value to competent teachers in appraising the work of individual students. The individual scores are of value to people outside the classroom, such as counselors, when they form a regular pattern over a number of years, and in occasional other circumstances, such as attempts to raise standards of promotion. Group scores can be of value to teachers in studying the relative progress of a class. Group scores, if they are interpreted and used with care, can be of value to others (principals, researchers, central office administrators, school boards, and the public) in making judgments about curriculum, programs, schools, and school systems.

COLLEGE ADMISSION TESTS

College admission tests differ from the standardized tests already discussed in two important respects. First, their use is less subject to the discretion of the schools. Schools have little choice but to offer college admission tests at the appropriate time to those who wish to take them. It is true that some of the tests discussed above, particularly some of the I.Q. and achievement tests, are not subject to the discretion of the local school or school system because they are required by the state school authorities or, in some cases, by Federal agencies in connection with Federally funded programs. But the college admission tests are *entirely* external in that not only their existence but their use is determined outside the schools. Secondly, in contrast to the tests already discussed, college admission tests are taken by only part of the age group to which they are addressed. Even though not all high school students take them, these college admission tests are important because of their role in guidance, in determining whether a young person goes to college, in deciding which college he goes to, and in granting advanced standing or college credit for work taken in high school.

The tests to be discussed in this group are the *Preliminary Scholastic Aptitude Test-National Merit Scholarship Qualifying Test* (PSAT-NMSQT), the *American College Test* (ACT), the *Scholastic Aptitude Test* (SAT), the College Entrance Examination Board's *Achievement Tests* (ACH), and CEEB's *Advanced Placement Examinations* (AP). The ACT is published by the American College Testing Program. The other tests are published by Educational Testing Service for the College Entrance Examination Board.

About a million high school juniors (about a third of the class) take the PSAT-NMSQT every year. Strictly speaking, it is not a college admission test, but it is related to college admission. The PSAT-NMSQT, as its name indicates, is a trial run for the SAT and is similar to it in all respects except length—two hours instead of SAT's three. It falls into two parts, verbal and mathematical, and is a general achievement test in these fields. It is entirely multiple-choice, with five choices for each question, but has a penalty for wrong answers (a correct answer counts one point in the raw score; a wrong answer counts minus one-quarter point; and a question not done, zero). The raw scores are converted to standard scores on a scale of 20 to 80,

which can be equated to the SAT score by adding a zero. For example, a 45 on the PSAT verbal is equal to a 450 on the SAT verbal. The PSAT goes back to 1959; in 1971 it was adopted as the NMSQT, replacing a different test previously used by the National Merit Scholarship Corporation.

In addition to giving the student an idea of how he will do on the SAT, the PSAT-NMSQT serves the purposes of the Merit Scholarship program. The "selection index" score used for Merit Scholarship purposes is the sum of twice the verbal PSAT score and the mathematics PSAT score; it ranges from 60 to 240. For example, a 65 on the PSAT verbal and a 70 on the mathematical would work out to 200 (twice 65, plus 70) as a "selection index" score. The top "selection index" scorers in each state (about one-half of one per cent of the secondary-school graduating class) are designated semi-finalists, and almost all of these are later confirmed as finalists by comparably high scores on the SAT and their schools' recommendations. Of these 15,000 finalists, about 3,000 will be awarded scholarships through the Merit program. But the semi-finalists who do not receive Merit Scholarships and an additional group of "commended" students (those in each state who have scores equal to or higher than the lowest state semi-finalist qualifying score but not high enough to qualify them as semi-finalists in their own states) have their achievement made known to colleges throughout the country. For students who make this charmed circle of semi-finalists and commended students, the test is a boost in college admission. For those who do not, the score is taken into account by high school counselors and others when considering college choice and whether the student should go to college at all.

About a million students a year take the ACT. A fairly new test, dating from 1959, it has been used mostly by public colleges and universities. A multiple-choice test that takes three and a half hours, it differs from the PSAT and SAT in three important respects. First, it has four parts: English usage, mathematics usage, social studies reading, and natural science reading. On each, the standard scores range from 1 to 36. Some colleges say that they find this four-way division of more value than the two-way division of the SAT. Secondly, there is no penalty for wrong answers. This means, on the one hand, that the test corresponds in this unfortunate regard to the other tests that the students have been taking at earlier points in their schooling but, on the other hand, that it is a less accurate measure of ability. Thirdly, the ACT usually cannot be taken a second time. In contrast,

the SAT can routinely be taken again (with a different set of questions, of course), thus reducing the importance of a single performance by the student.

Over a million students a year take the SAT, sometimes referred to as "the College Boards." The test dates back to 1926 in roughly its present form. It is very similar to the PSAT except that it is three hours long.[8] The standard scoring ranges from 200 to 800. For hundreds of thousands of students who hope to go to very selective colleges, this is the test that counts. It produces those two magic scores, one for the verbal test and one for the mathematical test, that mean so much to so many. It is well to remember, however, that except for very high and very low scores, there is no such thing as a "good" SAT score or a "bad" one except in terms of the admission standards of particular colleges. A 500-500 set of scores might be good for College A and poor for College B.

About 300,000 students take, in addition to the SAT, one or more of the CEEB Achievement Tests. These are one hour each in length and are offered in 14 fields: American history and social studies, biology, chemistry, English composition, European history and world cultures, French, German, Hebrew, Latin, literature, mathematics, physics, Russian, and Spanish. Some colleges require applicants to take from one to three of these tests and sometimes specify those that must be taken. For example, Kenyon College requires the English composition test and any other two; California Institute of Technology, mathematics and any two from biology, chemistry, English and physics; Yale University, any three. As their name indicates, these tests aim to measure the student's achievement in the particular subject. The tests are multiple-choice with a penalty for wrong answers. (Until 1971 there was a 250-word essay that constituted one-third of the English composition test score, but unfortunately it was abandoned, reportedly because of its expense.)

The CEEB's Advanced Placement Examinations (AP) are offered in 14 fields: English, music, mathematics, French language, French literature, chemistry, American history, classics, European history, German, physics, history of art, Spanish, and biology. Started in 1954,

[8] Beginning in the fall of 1974, the SAT will include a 30-minute, multiple-choice "Test of Standard Written English." The new test, being added because many colleges have complained of the lack of composition skills, will consist of questions on usage and sentence structure. To make time for the test, 15 minutes each will be cut from the verbal and mathematical sections. In a second change, subscores of the verbal test will be computed for reading comprehension and vocabulary.

they take up to three hours each. Their principal purpose is to give students an opportunity to earn advanced standing or advanced standing and credit from colleges for achieving college-level competence while still in high school. Advanced placement courses do not have to be taken in high school as a prerequisite to taking the tests, but they usually are. The tests are graded on a scale from 1 (low) to 5 (high), and usually a 3 or higher is necessary to gain advanced standing or advanced standing and credit. The decision is up to each of the 1,000 colleges that participate in the program, and some require a grade of 4 or higher for recognition. Strictly speaking, these are placement examinations rather than admission tests; no college requires any of them for admission. Every one of these examinations consists partly of essay or problem questions and partly of multiple-choice questions; they are unique in this respect among all of the standardized tests discussed in this paper. About 55,000 students wrote 71,000 AP examinations in 1973.

Many criticisms have been leveled at the college admission tests and their use. They have been accused of cultural bias. Critics have contended that the multiple-choice format fails to test writing ability or creativity and does not allow the student to explain his answers. It has been charged that too much weight is given to them in the admission decision. In recent years, some persons have said that the tests exist for the colleges rather than for the students.

These criticisms can be fairly judged only in light of the purposes and the record of these tests. The tests exist because more young people want to go to certain colleges than these colleges can admit. Selection could be made on the basis of any number of criteria, and in the past the most selective private colleges gave substantial weight to social class, wealth, religious affiliation, alumni connections, and attendance at a small number of private preparatory schools with which the colleges were familiar. When these colleges sought to broaden the sources of their students, it was impossible to know what grades and courses and class standing meant in the thousands of public and private secondary schools across the country. The college admission tests provided a means of judging the relative scholastic ability of students from these many schools. The tests opened up educational opportunities in these colleges to thousands of students from remote and little-known schools by enabling them to show that their ability was equal or superior to that of students coming from a small number of schools familiar to the admissions officers. Over the years the tests proved themselves by being good predictors of

academic success. In fact, they are usually the second-best single predictors of academic success as defined in terms of college grades. Only high school grades and class standing have a better record of prediction. But the test scores and high-school performance *taken together* constitute a better predictor than either alone, and the selective colleges take both into account.

Almost all of the major state universities and many of the other public colleges have been faced with a similar selection problem. Once they decided that they could no longer automatically admit all high school graduates from their areas who wanted to come, they had to select on some basis. They could admit purely on the basis of high school record, but this would ignore the fact that high school grades and courses are not uniform. Most public colleges, too, find admission tests of value in trying to select fairly.

Looking at the specific criticisms, the one of cultural bias is the most recent. As explained above, the admission test is a far less biased procedure than the one previously used by many private colleges. The admission of Jews, Catholics, blacks, and other minorities to some colleges was once governed by quota—if they were admitted at all. The increased use of the admission test generally coincided with an interest in increasing the variety of the students. It may be that the tests reflect some cultural bias, but since they serve as good predictors of college success by minority students, any cultural bias must also exist in the college program rather than just in the test itself. That is, the selective college may have a somewhat culturally biased program, but that is a separate issue from whether the admission test is culturally biased. Since the purpose of the admission test is to select the applicants who are most likely to succeed in certain colleges (not in life), if it does this well it serves its purpose. One factor that tends to offset any cultural bias peculiar to the test is the current practice of most selective colleges of accepting students from disadvantaged minorities with somewhat lower test scores and other qualifications than are required from other applicants.

The multiple-choice format of all the tests except the AP examinations is a shortcoming. As stated earlier, it does not test writing ability or creativity, nor does it allow a student to explain his answers, when such explanations might well show marked differences of ability among students.[9] But it would be considerably more expensive and

[9] The best-known critic of multiple-choice tests is Dr. Banesh Hoffmann, who criticizes them on these and other grounds. (See, for example, his *The Tyranny of Testing*, The Crowell-Collier Press, 1962.)

perhaps impossible to grade fairly a million test papers of the essay type. Moreover, the penalty for wrong answers that is a characteristic of the PSAT-NMSQT, SAT, and ACH tests eliminates one of the weaknesses of the typical multiple-choice test. Another practical advantage of the multiple-choice format is that it allows many more questions to be asked within a given time period.

The weight assigned to the admission test scores varies greatly among the colleges. To the best of my knowledge, no college has used or now uses test scores as the sole criterion of admission among candidates who get a certain minimum score. The usual practice is to sort applications into three piles as a first step: acceptable, doubtful, and unacceptable. It is easy to see that, when dealing with thousands of applications, the first screening has to be done by some kind of formula, and admission test scores certainly play a role in that formula. It is easy to see, too, how the public often gets the impression that a single point on the SAT score makes a life-or-death difference in the chances of admission to a given, selective college. But that single point can have significance only at the margin between the doubtful and the unacceptable piles in the first screening, and, even if the applicant had scored a point higher and found a place in the doubtful rather than the unacceptable pile, his chances of being eliminated on further examination would have been very great.

It is true that at the height of the admissions crunch several years ago, there was a tendency for some of the most selective colleges to let their cut-off scores drift upward. A very selective college might have put into the unacceptable pile any applicant with an SAT below 550 instead of using the previous cut-off point of 500. This was convenient for the colleges because it reduced the number of applications in the middle pile—the one that called for further, time-consuming, case-by-case consideration. Fortunately, there has been a reaction against this, and there has been a reduction, generally, in the weight that the scores receive. Two factors seem to be involved: a realization that the tests are not so accurate as to make a 50-point difference diagnostically important; and an appreciation that, at best, the tests measure only some of the characteristics that the colleges want in their students.[10]

[10] Incidentally, the SAT and most of the ACT scores have declined in recent years. The average SAT verbal score, for example, fell every year in the last ten, from 478 in 1962-63 to 443 in 1972-73. The average mathematical score also fell in every year, but somewhat less, from 502 to 481. This decline is not large, and the reasons for it are unclear. PSAT scores and ACT natural science scores have not declined.

Turning to the final criticism—that the whole admission test system serves the colleges rather than the students—one has to admit that the system helps the colleges if they want to select students on the basis of the likelihood of the students' succeeding at the particular institutions, because it permits the colleges to handle thousands of applicants from hundreds of secondary schools. But the admission test is of great value to the student, too, for only through the test can the student from the unfamiliar school "show his stuff." He may present a good high school record, which tends to demonstrate that he has exercised self-discipline, industry, and high aspiration over a period of years. But without the admission test, the college would have no way of knowing how his good performance at Crossroads High compares with the good performance of other students in hundreds of other schools. Another, perhaps even more important value of the admission test to the student is to help him choose a college of appropriate academic difficulty. Colleges and junior colleges in the United States present a great range in what they demand as standard performance. It is not in the young person's interest to select one too easy for him or one where he will be way over his head.[11] In the latter event, although transfer to a less demanding school is possible, it is inefficient, and a young student's unsuccessful experience at his first college must prove unpleasant at the very least, and could be traumatic. How much better to choose a school where one's chances of success are high. Since the use and refinement of the admission test procedure, the academic failure rate at many colleges has declined. This has been advantageous to the colleges, of course, but it has been even more advantageous to the young people involved, and they surely have even more at stake.

On the whole, college admission tests have a good record and continue to perform a useful function. There are a few easily remedied flaws, and it is true that the scores have not always been used reasonably. But those who would abolish them would be hard put to offer good alternatives.

---

[11] For example, in the fall of 1970 the average SAT scores for freshmen at Princeton University were 645 verbal, 674 mathematical; at Texas Christian University, 510 and 530; and at Ball State University, 420 and 430. Most of the Ball State freshmen would have flunked out at Princeton, and most of the Princeton freshmen had already achieved the work at Ball State. We can assume that each of the two institutions is a better choice than the other for its students.

SOME CURRENT CONTROVERSIES

Are the Tests Culturally Biased?

One of the recent criticisms levied against the standardized tests is that which I have mentioned in the previous sections, namely, that they are "culturally biased." The argument is that the tests ask for knowledge most familiar to the white middle class and are therefore biased against the poor, the blacks, the Spanish-Americans, and other minorities.

In order to discuss this criticism, cultural bias must be precisely defined. It is not enough to point out that the minority groups score lower, on the average, than does the white middle class. Bathroom scales are not biased against fat people because they record their obesity. Thermometers are not biased against sick people because they record their fevers. Stop watches are not biased against slow runners because they fully count the times they require to cover 100 yards. Cultural bias, to be a meaningful concept, must involve an *unfair* discrimination against a minority group. To determine whether a test unfairly discriminates, we must examine the purpose of the test and its use.

The group I.Q. tests, I said earlier, are culturally biased. The purpose of the tests is to determine fundamental learning ability, and in order to do this, the tests assume that all persons tested have had equal opportunity to acquire the knowledge and skills tested. Since they have not, the tests discriminate unfairly against the groups that have had below-average learning opportunities and discriminate unfairly in favor of groups that have had above-average learning opportunities. Group I.Q. tests were condemned in our earlier discussion, and cultural bias was one of the reasons.

The reading readiness tests are also culturally biased. The purpose of the tests is to measure certain knowledge and skills which are thought to be necessary before formal reading instruction can usefully begin. The pictures, words, and concepts used have a middle-class bias. But more important to the discussion than the tests themselves is the use to which they are put. On the basis of a single test which seems to confirm that most minority children are less ready to begin formal reading instruction—because they know less about standard English—most of the minority children are placed in slow groups or not allowed

to begin formal reading instruction at all. Thereupon they do not achieve as much as their "more ready" contemporaries—which is usually taken as proof that they were indeed not as ready to begin formal reading instruction! Reading readiness tests were rejected in our earlier discussion, and their cultural bias was one of the reasons for that rejection.

The elementary and secondary achievement tests present a mixed picture with regard to cultural bias. Some of the reading achievement tests, particularly in the early grades, *are* culturally biased. These tests purport to measure the achievement of elementary reading skill in standard English, and it is perfectly fitting that they do so because the achievement of such skill is a proper objective of our schools. But to the extent that the selection of vocabulary, pictures, and subject matter has a white middle-class bias, the tests tend to understate the reading skill of young minority children. I do not mean, however, that all minority children can read better than their reading scores indicate. As a matter of fact, many of them—the poorest readers—read worse than their scores indicate, but for reasons other than cultural bias of the tests. The achievement tests for the higher grades and high school are, for all practical purposes, not culturally biased. They test what they are supposed to test about as well for minority students as for majority students. To apply a special, lower standard to minority students would be an insidious, racist approach, as psychologist Kenneth B. Clark has pointed out.[12] Putting it somewhat differently, Stanley and Hopkins have said, "The fact that undernourished children weigh less than those who are well fed hardly builds a case to eliminate scales."[13]

As I have said earlier, the college admission tests serve their purpose of predicting college success about as well for minority students as they do for other students. Accordingly, they are not culturally biased as I have defined the term.

Is the Multiple-choice Format a Bad One?

With the exception of the Advanced Placement Examinations, all the standardized tests discussed are multiple-choice in format. Does

[12] In an interview published in the *Council for Basic Education Bulletin*, November 1969. Dr. Clark went on to say, "What advocates of such an approach are saying, if one takes it to its logical conclusion, is that these children are doomed, expendable in terms of not being capable of assimilation into the larger economic, educational, and cultural norms of the larger society."

[13] *Educational and Psychological Measurement and Evaluation* by Julian C. Stanley and Kenneth D. Hopkins (Prentice-Hall, 1972), p. 152.

the derisive epithet "multiple-guess" have any justification? I have touched on some of the criticisms in the preceding sections, but here I will elaborate on the disadvantages and advantages of the format.

Many who criticize multiple-choice tests because of the possibility of guessing do not understand the statistics involved. It is true that if a single problem, say the addition of 135 and 287, is given to a student as a straight computation problem, either he can do it or he cannot, whereas if he chooses the right answer from one of four in a multiple-choice test, there is no way on earth to tell whether he got the answer by skill or by a lucky guess. But if the student is given *six* such problems, his chances of getting them all correct by guessing falls to almost nothing—one in 4,096! If enough problems are given, then, the possibility that a high score could have been achieved by chance can be ruled out. But chance *can* play a role, and in score interpretation no achievement should be inferred from a score that may well be achieved by chance. For example, a score of 12 right out of 48 four-choice questions on a reading achievement test is the median chance score and must suggest the conclusion that the student cannot read at all. A score of 13, 14, or 15 probably warrants the same conclusion. Moreover, in interpreting any score, it should always be remembered that good luck may account for some of the correct answers. Accordingly, the student's "true score" may be somewhat lower.

Other criticisms of multiple-choice are more serious. The multiple-choice test calls only for recognition, not recall, and recognition involves a lesser grasp of knowledge. Trying to ascertain whether a student knows who was President during the Civil War might involve a multiple-choice question like this: "The President during the Civil War was: ( ) Washington ( ) Lincoln ( ) Hoover ( ) Jefferson." Or it could entail a direct question: "Who was President during the Civil War?" Some of the students who can choose (recognize) Lincoln from the four choices are not able to recall Lincoln in order to answer the latter question. Problems in non-school life normally require recall, not merely recognition. With more complicated questions, however, the score differences are probably not very great.

Another criticism of the multiple-choice format is that no creation or composition is required. This limitation means that no writing or organizing ability is demanded of the student, unlike most of the classroom examinations given in schools and colleges. It also means that the questions have to be narrower. Compare, for example, the multiple-choice question: "Which of the following is thought to be the commonest cause of inflation?" with the essay question: "Discuss

the causes of inflation." Perhaps it is here that one begins to appreciate the remark of the late President A. Whitney Griswold of Yale: "A few teachers still prefer written essays to intellectual bingo games that can be scored by electricity."[14] It is significant that every one of the Advanced Placement Examinations, given to determine whether students have mastered material equivalent to college-level courses, includes a number of non-multiple-choice questions.

A final criticism leveled at multiple-choice is that it penalizes the brilliant student, the deep thinker, the most knowledgeable. Dr. Banesh Hoffmann submits a number of examples in support of this contention. There are no doubt some faulty questions on the standardized multiple-choice tests. The important question here, however, is just how many there are, and how many students taking the tests are so penalized. An examination of the major tests does not show up many such ambiguous questions. Moreover, American students take multiple-choice tests frequently enough so that they "learn the ropes." To the extent that mastery of the multiple-choice technique involves a kind of mis-education, it is a disadvantage of the format. As Donald Barr has put it with respect to college entrance examinations, "the question is not whether the right people get into college: it is *what they must do to get into college* and *what happens to those who will not do it*."[15]

In view of the disadvantages of multiple-choice, are there offsetting advantages for standardized testing? The most obvious one, probably, is cost. Since large numbers of the tests are given, this is not an inconsequential matter. With 71,000 of the Advanced Placement Examinations to be graded, in part by human graders, the charge (in 1974) for each three-hour examination is $27.00. In contrast, the charge for the three-hour, machine-scored SAT is $6.50. And one must wonder where one would get enough graders to handle over a million SAT tests if they had to be graded with some kind of judgment. (They could hardly be graded locally because the grading standards must be made uniform in order for them to be fair to all concerned, whereas multiple-choice questions can be graded locally and by hand.)

There are other advantages. The standardized tests all involve a sampling of a large field of knowledge or skill. The fewer questions asked, the less valid the sample. Since essay questions take more time to answer, the multiple-choice format allows considerably more ques-

[14] In a speech given on January 29, 1952. Published in *Essays on Education* by Alfred Whitney Griswold (Yale University Press, 1954), p. 71.

[15] *Who Pushed Humpty Dumpty? Dilemmas in American Education Today* (Atheneum, 1971), p. 166. The emphasis is in the original.

tions to be asked in a given time period. Another advantage is that the multiple-choice answers are more subject to wide scrutiny and criticism than answers to essay questions would be. The reason for this is that the array of answers for a multiple-choice question must be determined in advance. (The specific questions and answers need, admittedly, more scrutiny and criticism from those outside the testing industry than they have been receiving.)

In summary, all things considered, the multiple-choice format seems to be a reasonable one for group standardized tests. Improvement in the questions is possible and sometimes needed. And exploration of alternative formats should be examined. For example, mathematics problems and many questions in the mechanics of writing could be "open-ended"; that is, simply require the student to figure out the answer. The answers would be simple enough to avoid the grading difficulties connected with essay questions.[16] Despite its advantages for standardized testing, the multiple-choice format should, as a general rule, not be used for non-standardized classroom testing. As a matter of fact, most good teachers do not use it in that way, except perhaps for an occasional quiz.

Are Criterion-referenced Tests Better?

Norm-referenced tests are so much the rule among standardized tests that many persons will have some difficulty imagining any other kind. Most of the standardized tests discussed earlier are *norm-referenced* tests: they report a student's score by comparing him with other students. This is true of the group I.Q. tests, the achievement tests, and all of the college admission tests except the Advanced Placement tests. But the reading readiness tests and the Advanced Placement tests are *criterion-referenced* tests: they report a student's score by comparison with learning objectives. In the reading readiness tests, the purpose is to ascertain whether the child has achieved certain knowledge and skills thought to be necessary before formal reading instruction is given to him. In the Advanced Placement tests, a certain score will qualify a student for advanced placement or advanced placement and credit in college, without regard to how many other students achieve such a score.

---

[16] Some of the alternative formats have been developed and successfully tested in the National Assessment of Educational Progress, and this experience is commented upon by the former head of that project, Frank B. Womer, in his article, "Reexamination of Multiple-choice Testing" (*Education*, April-May 1970).

If criterion-referenced tests tell us whether students have mastered instructional objectives, are they not better than norm-referenced tests, which tell us merely how well a student does in relation to other students? The answer depends on the purposes of the test.

The differences between norm-referenced tests and criterion-referenced tests are best understood by seeing how they are constructed. A norm-referenced test of third-grade reading, for example, begins with a study of what is common to the main textbooks used through third grade in present-day schools. This material is then sampled, usually by means of vocabulary and comprehension questions that differentiate between children. A question that almost all children would get right or one that almost all would get wrong is eliminated, no matter how important such a question is to reading achievement. The test results, then, will distribute the children along the familiar bell-shaped curve. That is the *purpose* of the test. A given child will get a score (grade-equivalent, percentile, stanine, percentile-band) that relates *his* achievement to the achievement of all children of the same age or grade placement in the nation. But except in this comparative way, it will reveal nothing about how well or how poorly he reads.

The criterion-referenced test begins differently. What, ask the test writers, *should* a child accomplish in reading by the end of the third grade? This question is answered in terms of specific objectives such as phonics knowledge, vocabulary, syllabication, correct tense forms, synonyms, and the meaning of common prepositions. Then questions are written in terms of these objectives. The second level of the *Prescriptive Reading Inventory*, published by California Test Bureau and one of the few criterion-referenced tests now available from a major test publisher, asks third-graders 153 questions related to 41 objectives. The test results purport to show how many and which of these objectives have been mastered by the student. Such information can be compiled for a class, a school, or a school system.

The norm-referenced tests, designed to compare the performance of an individual or a group with a large (usually national) group, tell nothing directly about achievement with respect to instructional objectives. The criterion-referenced tests, designed to compare the performance of an individual or a group with instructional objectives, tell nothing directly about how this performance compares with the performance of others. This fundamental difference suggests the reasons, in theory, for selecting one kind of test or the other. If the purpose is to compare individuals, schools, or school systems with

others, then the norm-referenced test is the better way to do it. If, on the other hand, the purpose is to find out just how well the individual, school, or school system has achieved certain instructional objectives, then the criterion-referenced test is the better way to do it.

The preference for criterion-referenced tests for certain purposes has to be stated in theoretical terms because test developers have only recently begun to show interest in criterion-referenced tests, and they have not been developed to the same extent as norm-referenced tests. As Dr. Ralph W. Tyler has put it, "the exercises [that make up norm-referenced standardized tests] have not been obtained by a systematic sampling of what children are expected to learn. Instead, the exercises comprise a sample of items that differentiate children. It may seem odd . . . but this is due to the fact that most psychometrists since World War I have been interested in individual differences among children and in the process of sorting children rather than in the processes of learning."[17]

Both norm-referenced tests and criterion-referenced tests aim at the same general objective: measurement of the achievement of individuals and groups. But the emphasis is different, as we have seen, and there is a growing interest in criterion-referenced tests. One reason is the current interest in accountability. It seems to many people more appropriate to measure school results by reference to instructional objectives rather than by comparisons with others. Another reason is the increased interest in individualized instruction. Criterion-referenced testing contemplates tailoring individual or class instruction to the objectives not yet mastered. Still another reason is the current antipathy to invidious comparisons. In norm-referenced testing, half the tested group will be below the norm ("below grade level") by definition (with reference to an up-to-date, accurate norm), and there is a tendency to regard this lower half as failures. They may not, in fact, have failed, if they have made steady progress toward worthy educational goals. The inner-city child who has learned to read, for example, has accomplished something very important and worthwhile, even if he does not read as well as the average child. (Similarly, the child with fine educational opportunities at home may read better than the average child and yet not have made as much progress as he could.) A fourth reason may be the feeling that criterion-referenced testing seems more in keeping with the proper aim of education: to have *every* student achieve as much as possible,

[17] *Council for Basic Education Bulletin*, March 1972.

without constant preoccupation with relative achievement. It may also be more conducive to the establishment of real standards of performance: minimum standards that almost all can hope to achieve, higher standards that some can hope to achieve.

Some criterion-referenced tests are already in widespread use in the schools. In the teaching of typing, for example, we can reasonably hope to instruct students so that almost all can achieve 50 words per minute with a minimum error rate. In the development of physical fitness, a standard test has been used for 15 years which has helped children and schools to achieve considerably higher levels of attainment. Outside the schools, perhaps the criterion-referenced test most widely in use is the driver's license examination. We would think it strange to have that test graded in terms of percentiles or age-level norms; we readily accept the notion that almost all persons can meet the minimum standards of the written and driving tests set as a requirement for obtaining a license. We would do well to study the implications of these examples for other subjects; that is what criterion-referenced testing is all about.

SUGGESTIONS FOR FURTHER READING

APPLEBEE, ARTHUR N.: "Silent Reading Tests: What Do They Measure?" *School Review*, November 1971. pp. 86-93. An account of a clever experiment which demonstrated some faults in some of the silent reading achievement tests.

CARVER, RONALD P.: "Reading Tests in 1970 Versus 1980: Psychometric Versus Edumetric," *The Reading Teacher*, December 1972, pp. 299-302. Contends that "edumetric" (criterion-referenced) reading achievement tests will replace "psychometric" (norm-referenced) ones, and that they should do so.

CHAUNCEY, HENRY, and JOHN E. DOBBIN: *Testing: Its Place in Education Today*, Harper & Row, 1963. Possibly the best-known and clearest explanation of standardized testing for laymen. Written by test makers, it is unduly favorable to the whole activity.

EBEL, ROBERT L.: *Essentials of Educational Measurement*, Prentice-Hall, 1972. An excellent college textbook. Covers teacher-devised tests as well as standardized tests. Some of the positions taken differ from those expressed in this paper.

FARR, ROGER, J. JAAP TUINMAN, and B. ELGIT BLANTON: "How to Make a Pile in Performance Contracting," *Phi Delta Kappan*, February 1972, pp. 367-69. A small experiment points up some of the hazards of using standardized achievement tests to measure the success of performance contracts.

FRY, EDWARD: "The Orangoutang Score," *The Reading Teacher*, January 1971, pp. 360-62. A satiric piece by the director of the reading center at Rutgers University which shows that "meaningful" grade-equivalent scores may well be obtained by guessing on well-known reading achievement tests. The three comments which follow and a rebuttal by Thomas F. Donlon (*The Reading Teacher*, October 1971, pp. 7-10) do not invalidate Fry's basic position.

GOSLIN, D. A.: "Standardized Ability Tests and Testing," *Science*, February 23, 1968, pp. 851-55. A discussion of criticisms of group I.Q. and achievement tests.

HAWES, GENE R.: *Educational Testing for the Millions: What Tests Really Mean for Your Child*, McGraw-Hill, 1964. A description of tests and common testing practices, written for parents. Chapter 1. "How Testing Shapes the Lives of Your Children," particularly recommended, even though it lacks any criticism.

STAKE, ROBERT E.: "Testing Hazards in Performance Contracting," *Phi Delta Kappan*, June 1971, pp. 583-89.

STANLEY, JULIAN C.: "Predicting College Success of the Educationally Disadvantaged," *Science*, February 19, 1971, pp. 640-47. A rather technical article defending the use of college admission test scores and high school grades for admitting "disadvantaged" applicants to selective colleges.

WOMER, FRANK B.: "Reexamination of Multiple-choice Testing," *Education*, April-May, 1970. A short article describing some of the limitations of the multiple-choice format and some of the alternative formats developed for the National Assessment of Educational Progress.

This is Number Twenty-two of a series of *Occasional Papers* published by the Council for Basic Education, 725 Fifteenth Street, N.W., Washington, D. C. 20005. These *Papers* are distributed without charge to all members of CBE at time of initial publication. Previous issues and additional copies may be purchased at the prices shown below. Orders for 20 or more copies of one title receive 20% discount. Other titles still in stock are:

No. 2: "What Every Intelligent Woman Should Know About Education" by Sterling M. McMurrin. 25¢.

No. 8: "Five Views of the 'New Math'" by Edwin E. Moise, Alexander Calandra, Robert B. Davis, Morris Kline and Harold M. Bacon. 25¢.

No. 10: "The Role of History in Today's Schools" by Erling M. Hunt, Charles G. Sellers, Harold E. Taussig and Arthur Bestor. 25¢.

No. 11: "Raising Standards in the Inner-City Schools" by Daniel U. Levine. 25¢.

No. 12: "The New English" by Sheridan Baker. 25¢.

No. 14: "The Electronic Revolution in the Classroom: Promise or Threat?" by Fred M. Hechinger, John Henry Martin and Louis B. Wright. 25¢.

No. 15: "What Are the Priorities for City Schools?" by Wilson C. Riles, Carl J. Dolce and Martin Mayer. 25¢.

No. 16: "Merit Pay and Alternatives: Descriptions of Some Current Programs" by George Weber and William H. Marmion. 25¢.

No. 17: "Phonics in Beginning Reading: A Guide for Teachers and Parents." 50¢.

No. 18: "Inner-City Children Can Be Taught to Read: Four Successful Schools" by George Weber. 50¢.

No. 19: "The New Case for Latin and the Classics" by John F. Latimer. 50¢.

No. 20: "What Is Career Education?" by Sidney P. Marland, Jr., James D. Koerner, and members of the CBE Board and staff. 50¢.

No. 21: "The Schools and the Press" by Gene I. Maeroff, Forbes Bottomly, Paul Woodring, and Barbara Reimers. 50¢.

BEST COPY AVAILABLE